Grid Technologies: Foundations for Preservation Environments

Presentations and discussions of how grid technologies can be used to support preservation environments

Moderator: Reagan Moore,

San Diego Supercomputer Center

Prototype Persistent Archive

- **Implementation validating preservation concepts**
  - Infrastructure independence
  - Separation of context management from content management
  - Preservation of authenticity and integrity
  - Submission pipeline
  - Technology management
- **Experience validating NARA digital holdings**
  - ARC metadata characterization (regular expression, XML schema, relational tables)
  - EAP collection preservation
- **Experience managing NARA digital holdings**
  - Import of record groups onto the Research Prototype Persistent Archive

Panel Members

- **Reagan Moore, San Diego Supercomputer Center**
  - Building preservation environments on data grids
- **(Ewa Deelman, University of Southern California)**
  - Automating processing with workflow environments
- **James D. Myers, Pacific Northwest National Laboratory**
  - Standardizing format descriptions
- **Geoffrey Fox, University of Indiana**
  - Portals for managing user interactions

Panel Session

- **Short presentations by panel members to introduce the topics**
- **Question and answer session with audience participation (7 panel questions)**

- **Goal is to identify the areas of most concern, and then discuss how grid technologies help provide an approach or solution**

Importance of Managing Authenticity and Integrity

- **Which area is of higher concern for long term preservation?**

1. **Mitigating against risk of data loss**

2. **Preserving authenticity of records**

Importance of Managing Growth of Archives

- **Which area is of higher concern for future preservation systems?**

1. **Scalability of processing environment for accession of records**

2. **Scalability of storage systems for managing hundreds of millions of records**

Importance of Managing Technology Evolution

- **Which area is at greater risk due to technology evolution?**

1. **Management of access mechanisms**

2. **Management of encoding formats**

Importance of Flexibility of User Interaction

- **Which area is higher priority for preservation environment interfaces?**

1. **Ability to manipulate preservation workflows**

2. **Ease of use of portal for interactive access**

Importance of Access Mechanisms

- **Which types of access are envisioned?**

1. **Bulk processing of archives contents to support knowledge discovery**

2. **Pervasive access through PDAs to discover single records**

Importance of Discovery Interface for the Archives

- **Which area is higher priority for describing archives content?**

1. **Unstructured information retrieval through text indexing**

2. **Structured information retrieval through queries on metadata**

Importance of Federation of Archives across Multiple Submitting Agencies

- **Which type of federation is of greatest interest?**

1. **Homogeneous environment with common preservation metadata and common encoding standards**

2. **Heterogeneous environment with multiple standards for encoding and descriptive metadata**

Building Preservation Environments on Data Grids

Reagan Moore

San Diego Supercomputer Center

Managing Distributed Data

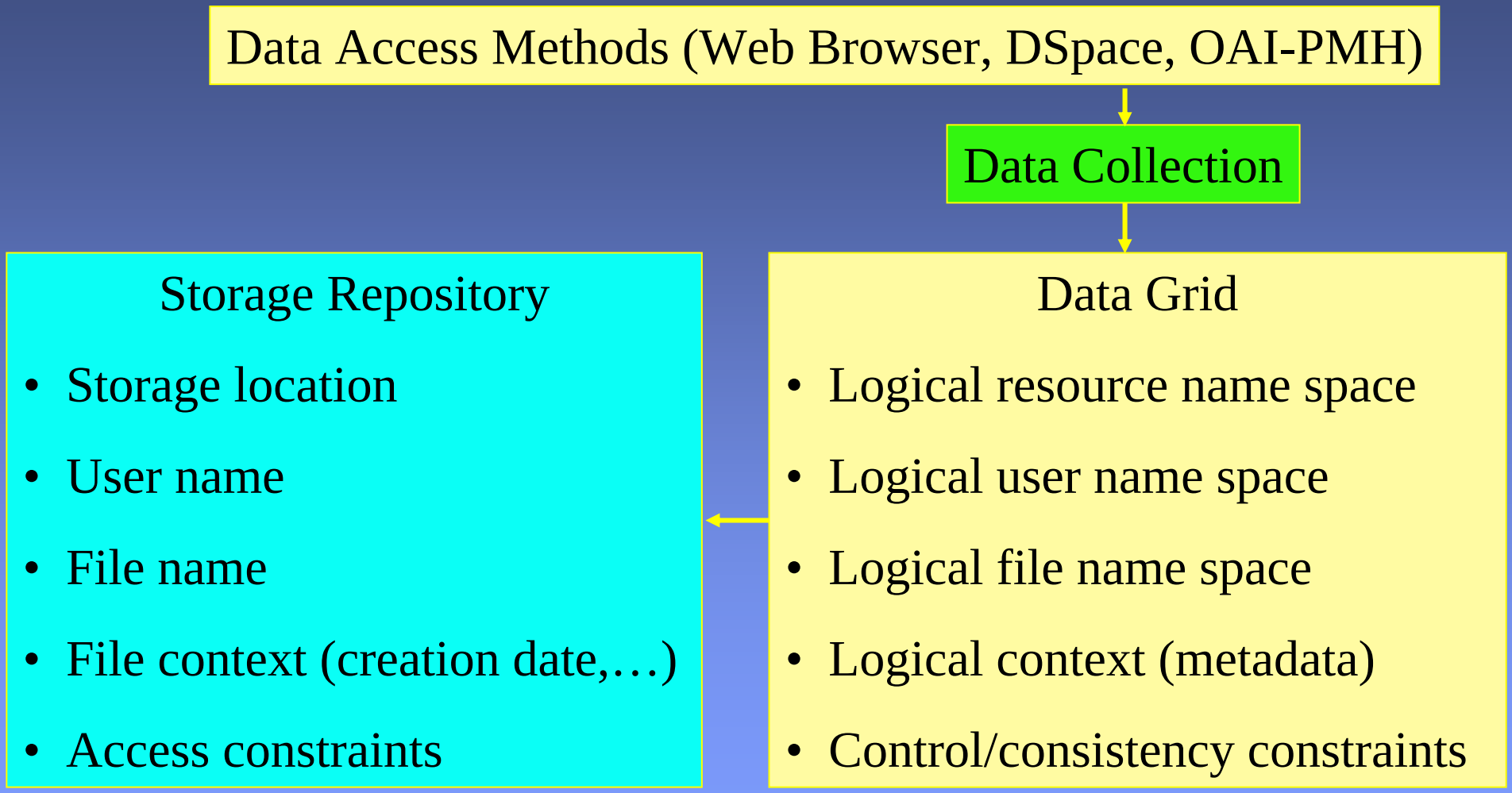
Data Access Methods (Web Browser, DSpace, OAI-PMH)

Storage Repository

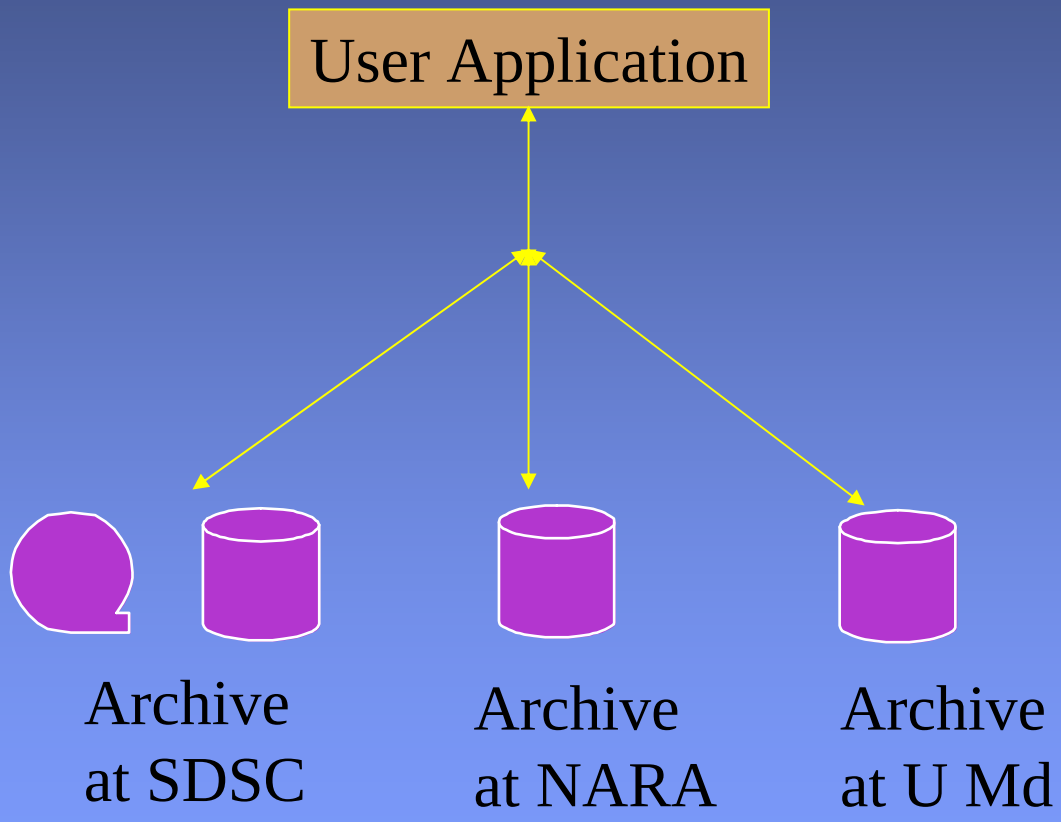
- Storage location
- User name
- File name
- File context (creation date,…)
- Access constraints

Naming conventions provided by storage systems

Data Grid

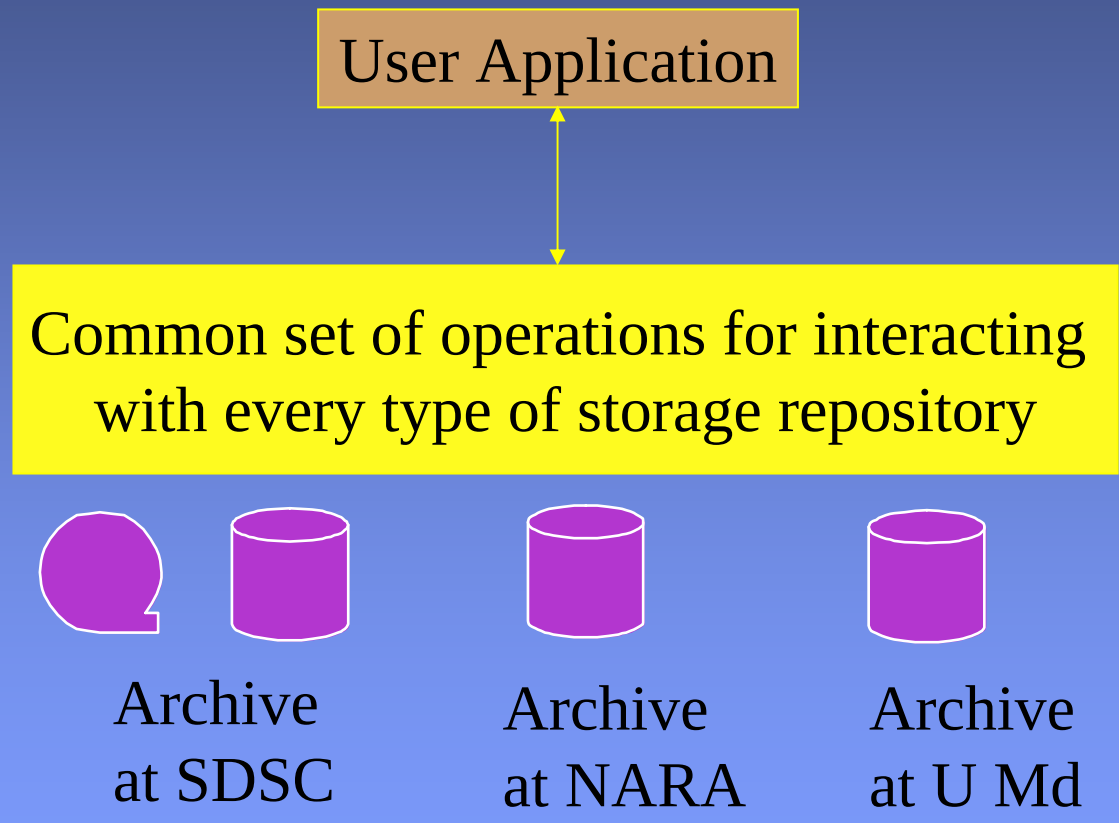

Data Access Methods (Web Browser, DSpace, OAI-PMH)
Data Collection
Storage Repository
• Storage location
• User name
• File name
• File context (creation date,…)
• Access constraints
Data Grid
• Logical resource name space
• Logical user name space
• Logical file name space
• Logical context (metadata)
• Control/consistency constraints

Accessing Multiple Types of Storage Systems



Standard Data Access Operations

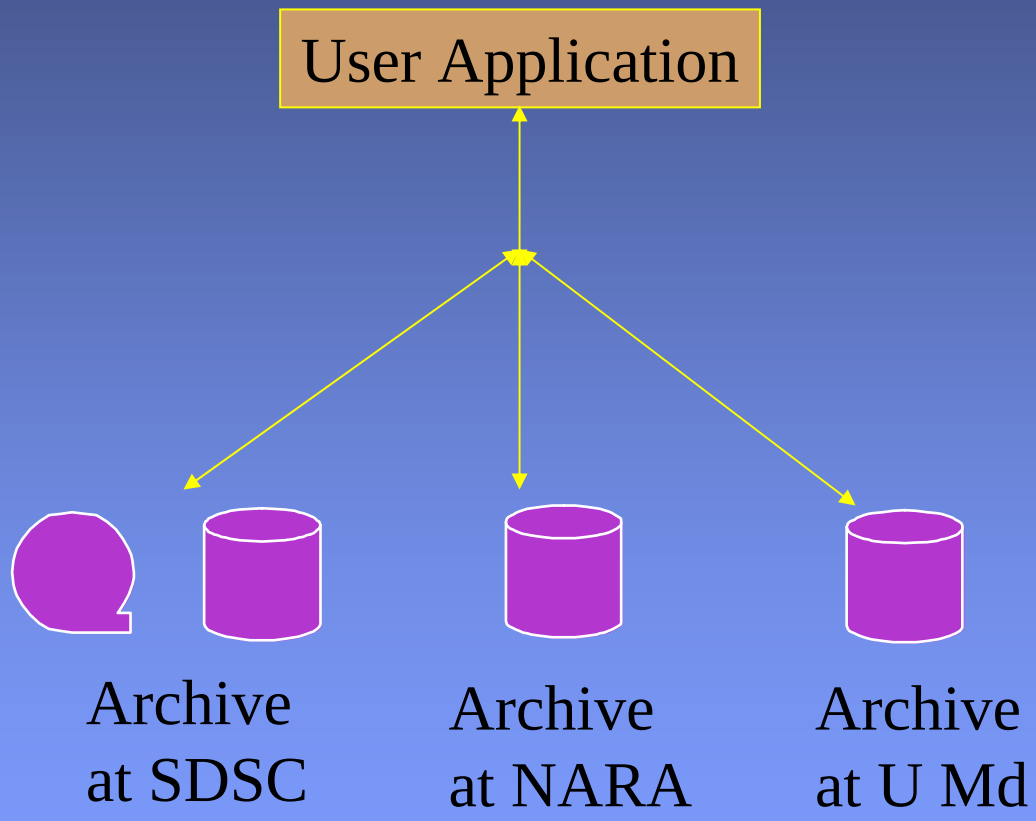
- Remote operations
  - Unix file system
  - Latency management
  - Procedures
  - Transformations
  - Third party transfer
  - Filtering
  - Queries
- Collective operations
  - Replication
  - Fault tolerance
  - Load leveling



Accessing Data at Multiple Sites

Each site has their own naming convention for files

A data grid provides a uniform way to name and access the files across the sites

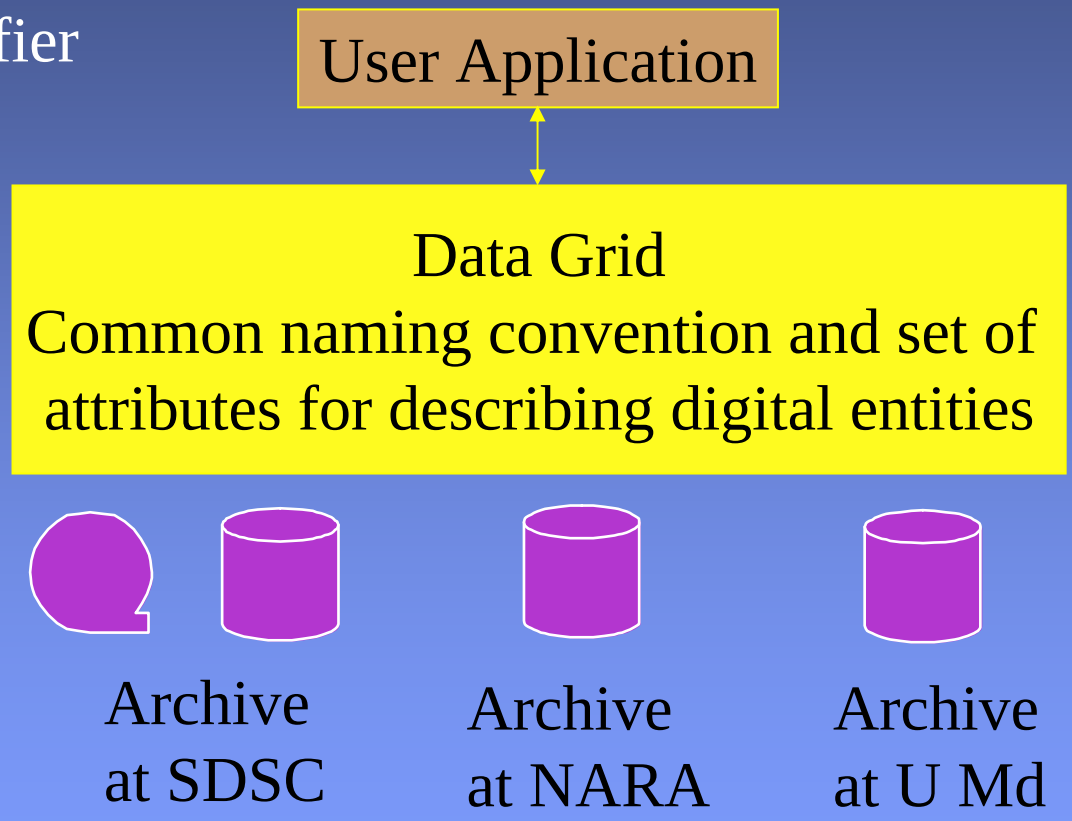


Building Distributed Collection

Logical name space
- Location independent identifier
- Persistent identifier

Collection owned data
- Authenticity metadata
- Access controls
- Audit trails
- Checksums
- Descriptive metadata

Inter-realm authentication
- Single sign-on system



Federation

Data Access Methods (Web Browser, DSpace, OAI-PMH)

Data Collection A

Data Collection B

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Access controls and consistency constraints
on cross registration of digital entities